CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Institutional Knowledge at Singapore Management University

Singapore Management University
Institutional Knowledge at Singapore Management University

Research Collection School Of Information Systems | School of Information Systems

10-2009

Semantics-preserving bag-of-words models for efficient image annotation

Lei WU

Steven C. H. HOI
Singapore Management University, CHHOI@smu.edu.sg

Nenghai YU

DOI: https://doi.org/10.1145/1631058.1631064

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research


Part of the Databases and Information Systems Commons, and the Data Storage Systems Commons

Citation

WU, Lei; HOI, Steven C. H.; and YU, Nenghai. Semantics-preserving bag-of-words models for efficient image annotation. (2009). *LS-MMRM '09: Proceedings of the First ACM workshop on Large-scale Multimedia Retrieval and Mining, Beijing, October 23*. 19-26. Research Collection School Of Information Systems.
Available at: https://ink.library.smu.edu.sg/sis_research/4189

This Conference Proceeding Article is brought to you for free and open access by the School of Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email libIR@smu.edu.sg.

Semantics-Preserving Bag-of-Words Models for Efficient Image Annotation

Lei Wu *
MOE-MS Keynote Lab of MCC
University of Science and Technology of China
Hefei, China 230026
leiwu@live.com

Steven C.H. Hoi
School of Computer Engineering
Nanyang Technological University
Singapore 639798
chhoi@ntu.edu.sg

Nenghai Yu
MOE-MS Keynote Lab of MCC
University of Science and Technology of China
Hefei, China 230026
ynh@ustc.edu.cn

ABSTRACT

The Bag-of-Words (BoW) model is a promising image representation for annotation. One critical limitation of existing BoW models is the semantic loss during the codebook generation process, in which BoW simply clusters visual words in Euclidian space. However, distance between two visual words in Euclidean space does not necessarily reflect the semantic distance between the two concepts, due to the semantic gap between low-level features and high-level semantics. In this paper, we propose a novel scheme for learning a codebook such that semantically related features will be mapped to the same visual word. In particular, we consider the distance between semantically identical features as a measurement of the semantic gap, and attempt to learn an optimized codebook by minimizing this gap. We refer to such a new codebook method as Semantics-Preserving Codebook (SPC) and the corresponding model as Semantics-Preserving Bag-of-Words model (SPBoW). This novel model generates codebook for each object category and only needs to update the codebook for a specific category when incomes an object, which makes it convenient to scale up with the increasing number of objects. Experiments on image annotation tasks with a public testbed from MIT's *Labelme* project, which contains 11,281 objects of 495 categories, show that the SPC learning scheme is efficient in handling large number of objects and is able to greatly improve the performance of the existing BoW model.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; I.2.6 [**Artificial Intelligence**]: Learning

*This work was performed when Mr. Lei Wu was a research assistant at Nanyang Technological University.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
LS-MMRM'09, October 23, 2009, Beijing, China.
Copyright 2009 ACM 978-1-60558-756-1/09/10 ...$5.00.

General Terms

Algorithm, Experimentation

Keywords

Image annotation, bag-of-words model, semantic gap, distance metric learning

1. INTRODUCTION

With the popularity of digital cameras and high quality mobile phones, huge number of novel objects appear in Web photos. Massive novel objects have posed a great challenge for image retrieval tasks. Automatic image annotation is one promising solution to address this challenge. Generally, automatic image annotation is the process of employing computer programs to automatically assign an unlabeled image a set of keywords or tags, each of which represents some semantic object/concept. By the auto-annotations, an image retrieval problem is turned into a text retrieval task, which can be effectively resolved by taking advantages of mature text indexing and retrieval techniques. Image annotation has become one of most active and important research topics in multimedia research.

In the past decade, there have been numerous studies on automatic image annotation [20, 10, 5, 9]. Some earlier studies often extract global visual features, such as color and texture, from whole images to represent images as data points in a vector space. As a result, the image annotation task is turned into a supervised classification problem with data on some vector space [5]. Such an approach enjoys the merits of efficient computation and compact storage, but often only works effectively for annotating scene images or single-object images. They usually performed poorly for general images that contain multiple objects.

Later, instead of simply extracting global features from whole images, more promising studies have been focused on regional features. One typical approach is to partition an image into multiple regions/blobs based on image segmentation and clustering techniques. As a result, image annotation can be turned into a machine translation task of classifying regions/blobs into keywords [9]. Along this direction, a variety of statistical learning techniques have been applied to model the relationships of words and regions [38]. The performance of these approaches sometimes depend on the quality of image segmentation, which is still a research challenge in image processing.

Recently, thanks to the advances of powerful local feature descriptors, such as SIFT [27], researchers in computer vision have attempted to resolve object recognition/image annotation problems using a new approach, known as "bag-of-words" (BoW) model, which was derived from natural language processing. Specifically, given an image, BoW first employs some interest point detector, e.g. the DoG (Difference of Gaussians) detector, to detect salient patches/regions in the image. Further, some feature descriptor, e.g. SIFT, is applied to represent the local patches/regions as numerical vectors. The last step of BoW is to produce a codebook by converting vectors representing the patches to "codewords", e.g. applying k-means clustering on all the feature vectors and defining the codewords based on the centers of the resulting clusters. By mapping each patch in the image to a certain codeword, the image can thus be represented by the histogram of the codewords. Based on the BoW representation, some well-known topic models, such as probabilistic latent semantic analysis (pLSA) [15] and latent dirichlet allocation (LDA)[3], can be applied to analyze the topics of the images [7]. Sacrificing the spatial information, the BoW model is quite efficient on large datasets, and enjoys promising performance on object categorization [23].

However, there are several notorious drawbacks with BoW. Besides the ignorance of spatial information, an issue that has been discussed in many recent papers [22, 25, 4], another critical disadvantage is that semantics of the objects is considerably lost during the detection of sub-regions and the generation of visual words in the codebook generation process. Firstly, the detection and segmentation of sub-regions damages the semantic integration. Several methods have been proposed to locate the sub-regions in the image, e.g. regular grid [38], interest point detector [33, 26], random sampling [28], sliding windows [21], other segmentation methods [14] etc. However, due to the lack of human knowledge, these methods cannot locate the semantically intact regions very accurately, which partially causes the semantic gap problem. Secondly, it is problematic for generating the visual words using k-means clustering in Euclidian space, which implicitly assumes that SIFT features of similar semantics are distributed in the same clusters in Euclidian space. This however does not always hold, especially for high dimensional SIFT features. Rong et al. [41] proposed to learn a unified visual bits for object recognition which works well on small number (around ten) of objects. Unlike the completely unsupervised clustering by k-means in visual word generation or supervised combination of visual bits, we believe that a semi-supervised clustering approach with the aid of side information could lead to more effective codebooks for representing objects towards object categorization and annotation tasks.

To this end, this paper proposes a novel **semantics preserving bag of words** model (SPBoW), which considers the distance between the semantically identical features as a measurement of the semantic gap, and tries to learn a codebook by minimizing this semantic gap. We formulate the codebook learning task as a distance metric learning problem which can be expressed as a quadratic program (QP). We then propose an efficient eigen-projection algorithm to solve the optimization problem efficiently. With the integrated knowledge and side information, the semantic gap can be minimized and the codebook is able to consistently represent the semantic of the objects. Besides, this approach generates codebook for each object category and for a novel category the method only needs to update the codes for that category, which make it very efficient in handling large number of objects. Experiment on 11,281 objects from 495 categories demonstrates the efficiency and effectiveness of the proposed method in handling large scale objects. Besides, to the best of our knowledge, this is the first metric learning approach to overcome the semantics lost limitation of regular BoW models.

As a summary, the main contributions of this paper include: (1) we propose a measurement of the semantic gap in the codebook generation process; (2) we suggest a novel approach to learn a codebook by minimizing the semantic gap; (3) we propose and implement an efficient algorithm to solve the codebook learning task; and (4) we evaluate and compare a number of different methods for the codebook generation process in building various bag-of-word models towards object annotation tasks.

The rest of the paper is organized as follows. Section 2 reviews some of the related work. Section 3 discusses the framework of the SPBoW model. Section 3.2 gives the details of the object representations for this novel model. Section 4 elaborates on the codebook learning task and formulates the task as an optimization problem. Section 4.2 discusses the solving of the optimization. Section 5 applies the learnt codebook on object annotation tasks. Section 6 compares the SPBoW model and the metric learning algorithm with several state-of-the-art methods for object annotation experiments on the Labelme testbed [29]. Section 7 concludes the paper.

2. RELATED WORK

Our work is related to several research topics, including image annotation and object recognition, and distance metric learning. Below we briefly review the related work of both categories, respectively.

2.1 Image Annotation and Object Recognition

In literature, numerous studies have been devoted to image annotation and object recognition. They can be roughly grouped in three major categories. The first category is based on global features extracted from whole images [12]. As a result, regular supervised classification techniques, such as SVM, can be applied to solve the categorization and annotation tasks.

The second category is to extract regional features such that an image can be represented by a set of visual regions/blobs [2, 20]. The image annotation task is thus converted to a problem of learning keywords/tags from visual regions/blobs. For instance, Barnard et al. [2] treated image annotation as a machine translation problem. Jeon et al. [9] proposed the cross-media relevance models (CMRM) model, which combines both surrounding texts and image contents for annotation. Jin et al. [20] studied coherent language models that takes into account the word-to-word correlation.

The last category is focused on applying bag-of-features or bag-of-words representations for image annotation/object recognition tasks [6, 32]. Csurka et al. [6] proposed a bag-of-keypoints approach similar to BoW in text categorization for visual object categorization. Jiang et al. [19] studied some practical techniques to improve the performance of bag-of-features representation for object recognition and retrieval.

Tirilly et al. [32] proposed a language modeling approach to address one limitation of bag-of-word models, i.e., the loss of spatial information.

In addition, there are also some emerging paradigms for image annotations, such as search-based annotation [34, 37] that explores WWW images in helping the annotation tasks, and the ALIPR paradigm proposed in [24], which used advanced statistical learning techniques to provide fully automatic and real-time annotation for digital pictures. These techniques are not very relevant to our focus, and thus out of further discussions.

2.2 Distance Metric Learning

From a machine learning point of view, our work is related to distance metric learning (DML). Specifically, consider a set of n data examples $X = \{x_i \in \mathbb{R}^d\}_{i=1}^n$ in d-dimensional vector space, the objective of DML is to find an optimal Mahalanobis metric M from training data (side information) that can be either class labels or general pairwise constraints [39]. The class labels can be obtained by manual labeling or collected from user-generated tags from WWW. The pairwise constraints could be obtained from some multimedia systems, e.g. the relevance feedback logs from CBIR [16, 30, 18]. In literature, DML has been actively studied in recent years. Existing DML studies can be roughly grouped into two major categories. One is to learn metrics with class labels, such as Neighbourhood Components Analysis (NCA) [13], which are often studied for classification [11, 35, 40]. The other is to learn metrics from pairwise constraints that are mainly used for clustering and retrieval. Examples include RCA [1] and Discriminative Component Analysis [17], amongst others [39].

3. FRAMEWORK FOR SEMANTICS PRESERVING BAG-OF-WORDS MODELS

3.1 Overview

The BoW model treats an image as a bag of "codewords", which consists of lots of independent local appearance features. These features are either located by salient region detectors like SIFT, random samplings like random windows, segmentation, or regular grid. Since the high dimensional feature may contain much noise and redundancy, and it is difficult to store and use them, visual words are further generated by clustering these features. Each visual word in fact corresponds to a region in the feature space. All the visual words fully divided the feature space into many sub-regions. With these visual words, the model is more robust and efficient. Ignoring the spatial dependency, which may make the model too complex for large scale data, this method models the objects by their visual words histograms over the images.

The basic idea of applying the BoW model on object annotation is to apply the model to a Naïve Bayes classifier [36] or more complex latent topic analysis methods such as pLSA [31], LDA [3]. For the Naïve Bayes case, object annotation equals to matching the histogram visual words in an image with the histogram of visual words of each object. If the histogram matches certain object, the name of the object is annotated to the image. For the pLSA and LDA methods, which are usually adopted in an unsupervised topic detection task, objects are taken as the hidden topics of the images. The object category is unknown and large computational cost are introduced by the latent topic analysis, which are not suitable for large scale data.

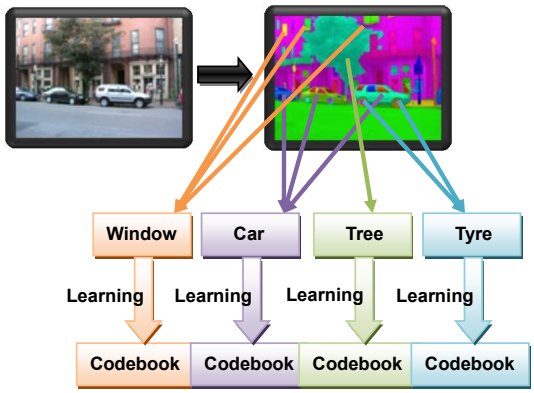


Figure 1: Flowchart illustrating the process of building the semantics-preserving bag-of-words model.

In this paper, we adopt the improved bag of words model with the Naïve Bayes framework for the large scale web image annotation task.

The flowchart of the proposed method is shown in Fig. 1. In the training process, all the objects in the images are segmented and tagged by the users. SIFT features are adopted to represent the local appearance of these objects. The SIFT features which are located at the similar part of the same object are considered relevant and put into the same chunklet (bag). With this side information provided, supervised metric learning is applied to train a proper distance metric. This distance metric is adopted to cluster these appearance features for each object category. With these codebooks, we can estimate the existence of these objects in a Naïve Bayes way.

3.2 SPBoW for Object Representation

In the traditional bag of words model, an object is represented by the histogram of the visual words in the image, where the object exists. This simple representation has some serious drawbacks. First of all, both the words for the objects and the words for the background are all incorporated into one model. Since this representation does not separate the objects from the background, it brings background noise into the model which is supposed to capture the object. Secondly, this representation can only deal with the single object case. In most cases there are multiple different objects in one image and if the modeling process does not separate them, all these objects will become noise to each other in the modeling process. Although the latent topic analysis can somewhat deal with this problem, it brings another challenging problem, how to determine the number of latent topics. Thirdly, this representation only captures one instance of the object which are shown in the image. Generally speaking, each object may have various transformations and perspective alternations. Representation based on one image loses much information about the object.

For the above reasons, we represent the semantics by modeling each individual object rather than modeling each image. We adopt image data from MIT's Labelme testbed [29], since objects in this dataset are segmented and labeled in each of the images by the users. Then the SIFT descriptors

are generated for each of the images. The SIFT features which are located on the same object in all the images are collected to represent this object.

In order to preserve the semantics in the bag of words model, all the features of the object are clustered into one or several discriminative visual words, which are only used to represent the specific object. The visual words for an object may denote different parts of the object or different views of the object.

4. LEARNING TO OPTIMIZE CODEBOOK

The codebook generation process is critical to building the BoW model. Instead of generating the codebook by a completely unsupervised learning approach (typically k-means clustering with Euclidean distance) that often leads to much loss of semantic information, in this paper, we suggest a novel learning scheme by exploiting side information for optimizing the codebook generation process.

4.1 Problem Formulation

Assume we are given a set of feature instances $X = \{x_i\}_{i=1}^N$ and instance constraints $Z = \{z_i\}_{i=1}^N$,where z_i indicates whether or not the data instance x_i is on the object or not. Each object in a training image is segmented into semantic parts and labeled manually. If the feature point is inside any labeled segments, $z_i = 1$; otherwise $z_i = 0$. Also there are a set of pairs $\{(x_i, x_j)|x_i, x_j \in X\}$, and corresponding pair-wise constraints $Y = \{y_{ij}\}$, which indicates whether the pair (x_i, x_j) is located in the segment with same label or not.

So this task tried to learn a similarity metric A to measure the distance between these visual features.

$$d(x_i, x_j) = \sqrt{(x_i - x_j)^\top A(x_i - x_j)} \tag{1}$$

where x_i and x_j are two sample points, and A is the learnt metric which is positive and semi-definite w.r.t. the property of a metric.

This task can be formulated as a distance metric learning problem where the overall distance between features of the same objects should be minimized. In particular, we formulate the problem as an optimization task as follows:

$$\min_{A\succeq 0, b} \quad \sum_i z_{i1}z_{i2}\xi_i + \frac{\lambda}{2}tr(AA^\top) \tag{2}$$

$$s.t. \quad y_i(\|x_{i1} - x_{i2}\|_A - b) \leq \xi_i, i = 1, \ldots, N \tag{3}$$

$$\|A\| = 1/\sqrt{\lambda} \tag{4}$$

where $A \in \mathbb{R}^{d\times d}$, x_{i1} and x_{i2} are two SIFT features. $z_{i1}, z_{i2} \in \{0, 1\}$ indicate whether the features are located on the semantic object in the image. If x_{i1} is located on the object, $z_{i1} = 1$; otherwise $z_{i1} = 0$. $y_i \in \{-1, 0, 1\}$ represents the two features are of similar semantic. $y_i = 1$ represents the two features are located at the same semantics; $y_i = -1$ represents the two features are not at all related; $y_i = 0$ if not decided or no human knowledge provided.

The learning task is to minimize the loss that the features which are located at the same semantic part of the objects are assigned to different codes and the loss that two different features are mapped to the same code. The first inequality condition is to prevent the semantic gap from being too large. The second inequality condition ensures the learned metric is positive definite and does not become 0. z_{i1} z_{i2} ensure the training features are all located on the meaningful objects. Solving the optimization problem, we can get the distance metric A and the threshold b. A indicates the proper distance metric, under which the distance between features is measured, and b is the threshold to decide whether two features should be considered identical or not.

4.2 Optimization

The above optimization generally belongs to an semi-definite programming (SDP) problem, which is often difficult to solve for large scale applications. To address this challenge, we propose an efficient iterative algorithm based on gradient descent techniques. In our algorithm, subsets $\mathcal{S}_t^+$ and $\mathcal{S}_t^-$ are firstly defined. $\mathcal{S}_t^+$ contains all the feature pairs that should have the same semantic while long distance under current metric. $\mathcal{S}_t^-$ contains all the feature pairs that should have the different semantic while short distance under current metric. These pairs of features are used as the training data. The output of the algorithm is a $d \times d$ metric A and the threshold b.

In the iteration process, firstly we define an learning rate. This rate decreases with the iteration to make sure of the fast convergence. Secondly, we update the subset for training. Thirdly, the gradient w.r.t the learning metric A and threshold b are calculated. Fourthly, we update the metric A and the threshold b, and finally we project all the eigenvalue of the metric A to positive to ensure its metric property. This learning process converges when the training subset is empty or metric A does not alter. The whole algorithm for learning the semantics preserving metric is listed as follows.

Algorithm 1 Semantics-Preserving Metric Learning

1: INPUT:
- SIFT feature matrix: $X \in \mathbb{R}^{N\times d}$
- pair-wise constraint $(x_{i1}, x_{i2}, z_{i1}, z_{i2}, y_i)$, where x_i is the i^{th} SIFT feature, z_i indicate whether the location of the feature has semantic meaning, and constraints $y_i = \{+1, 0, -1\}$ represents the two features are of the same semantic, not known, or different semantic.
- regularization parameter λ
- learning step size parameter γ

2: OUTPUT:
- feature matric A, code threshold b

3: initialize metric and threshold: $\alpha = I$, $b = b_0$
4: set iteration step t = 1;
5: **repeat**
6: (1) update the learning rate:
$\gamma = \gamma/t$, t = t + 1
7: (2) update subset of training instances:
$\mathcal{S}_t^+ = \{(x_{i1}, x_{i2}, y_i)|(1 + y_i)z_{i1}z_{i2}\|x_{i1} - x_{i2}\|_A^2 < 1\}$
$\mathcal{S}_t^- = \{(x_{i1}, x_{i2}, y_i)|(1 - y_i)z_{i1}z_{i2}\|x_{i1} - x_{i2}\|_A^2 > 1\}$
8: (3) compute the gradients w.r.t A
$\nabla_A\mathcal{L} = Z_1Z_2(\lambda A + D_X^\top Y^\top D_X$,
$D_X = X_1 - X_2$,
$X_k = [x_{1k}, x_{2k}, \cdots, x_{nk}]^\top$,
$Z_k = diag(z_{1k}, z_{2k}, \cdots, z_{nk})$
9: (4) compute the gradients w.r.t b
$\nabla_b\mathcal{L} = tr(Z_1Z_2Y)$
10: (5) update metric and threshold:
$A = A - \frac{\gamma}{t}\nabla_A\mathcal{L}$, $\quad b = b - \frac{\gamma}{t}\nabla_b\mathcal{L}$
11: (6) project A back to PSD cone:
$A = K'\Lambda K$
$\hat{\Lambda} = max(0, \Lambda)$
$A = K'\hat{\Lambda}K$
$A = \frac{1/\sqrt{\lambda}}{\|A\|}A$
12: **until** convergence

4.3 Codebook Generation

A codebook can be generated by clustering the features under the learned distance metric into some visual words or codes. Different visual words should represent different views or different parts of an object. In this paper, we generate the codebook for each object category, so that the linkage between the codes and high level semantic of object category can be obtained. This linkage serves to connect the low level feature to high level semantic objects.

For each category, we collect all the related features, and k-means clustering based on the learnt metric is performed to generate some clusters (codes) for this category. The codes for all categories form a global codebook, where each code only corresponds to one object. Thus this codebook is called semantics preserving codebook (SPC).

One of the problems with SPC is to decide the number of codes for each category. In this paper, we measure the visual complexity by the information theory. Consider each category as a bag. Each feature is generated from the bag with certain probability, which can be estimated by either the distance to average feature vector or the frequency of the features. Based on this probability, we calculate the information entropy of the bag as the visual complexity. If the features in a bag are similar, the entropy is small; otherwise entropy is large. Finally, the number of codes assigned to the object is proportional to its visual complexity.

$$p(x_i|c) = \frac{1}{\sqrt{2\pi}\sigma} \exp^{-\frac{\|x_i - \hat{x}\|_A^2}{2\sigma^2}} \tag{5}$$

where c is an object category, $\hat{x} = \frac{1}{n_c}\sum_i^n x_i$, and n_c is the total number of features related to the object. $p(x_i|c)$ is the estimated probability of x_i generated from object c. The visual complexity of the object $H(c)$ and the size of codebook assigned to object $L(c)$ are calculated as follows.

$$H(c) = -\sum_{i=1}^{n} p(x_i|c) \log p(x_i|c) \tag{6}$$

$$L(c) = L_{max} \times \frac{H(c)}{\log n_c} \tag{7}$$

where L_{max} is the maximum size of the codebook for each object.

With the assigned codebook size for each object, we can generate the codebooks by k-means clustering. The input of the clustering algorithm is the collections of features related to each object and the global metric A. The output is the clusters (codes) $\{w_{ij}|w_{ij} \in c_i\}$ for each object c_i and the range r_{ij} for each code. Here we use the maximum distance of feature to the code's center as the range of the code.

$$r_{ij} = max_{x \in w_{ij}} \|x - w_{ij}\|_A \tag{8}$$

In the assignment of visual word, if a feature is inside the range of any code, the code is assigned to the feature, otherwise the feature does not correspond to any code and is discard. This is for the consideration of the unknown features and noise. Suppose x is the feature, w_{ij} is the j^{th} cluster (code) for object c_i.

$$w(x) = \begin{cases} w_{ij}, & \delta(\min_{ij} \|x - w_{ij}\|_A, r_{ij}) = 1; \\ discard, & \text{otherwise.} \end{cases} \tag{9}$$

$$\delta(x, b) = \begin{cases} 1, & x < b; \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

For this visual word assignment, each feature may be assigned to multiple codes if it lies in their ranges.

5. OBJECT ANNOTATION USING SPBOW

Using the semantics-preserving codebook (SPC) and the SPBoW representation, object/image annotation can be easily performed in a Naïve Bayes classifier.

Given a novel image, SIFT features are firstly extracted. Then these features are assigned to visual words (codes) if they lie in the range of these visual words.

$$Coding(x) = \begin{cases} c_i, & \|x - c_i\|_A < r_i, i = 1 \cdots K; \\ discard, & \text{otherwise.} \end{cases} \tag{11}$$

Different from the traditional bag of words method, in this paper, the visual features can be assigned to multiple visual words in different object categories, since the range of the visual words may overlap each other. Here we assign the visual words by range r_i rather than assigning the feature to the closest visual word because the range of words from different object categories may overlap each other, and the feature may be assigned to multiple visual words. This makes sense since the same semantics may appear in different objects. For example, the "window" can appear in both building and car. Assigning the visual words by their ranges can better handle this case than assigning these visual words to the closest cluster.

Thirdly, histograms of the visual words are generated for each object.

$$histogram(i) = f(c_i) / \sum_j f(c_j) \tag{12}$$

Then the tags are ranked by the frequency of their related visual words. The top N tags are used to annotate the image. In this paper, we have evaluated the performance under $N = 1, \cdots, 10$.

6. EXPERIMENTS

In this section, we evaluate the semantics-preserving bag-of-words (SPBoW) model with the traditional bag-of-words (BoW) model for image annotation. We evaluate the proposed models in several aspects. In addition, our SPBoW framework can also be integrated with other existing distance metric learning techniques. In our experiments, we also evaluate different implementations of SPBoW by adapting other existing DML algorithms in our framework. We are interested in examining their empirical performance.

6.1 Experimental testbed

We employ a dataset from Labelme project [29], which consists of around 11,281 objects of 495 categories related to downtown streets. The objects include cars, trees, buildings, persons, lights, ladders, sidewalks,air conditions, mail box, signs, bicycle, umbrella, etc. There are more than 400,000 local appearance features in this dataset for learning the codebook. The experiment will show the efficiency of the proposed approach on this dataset, and It is worth noting that in case of novel incoming objects, the approach does not need to re-build the codebook of former objects and

only need to combine the codes for novel objects into the codebook. Thus this approach is capable to scale up with increasing of objects easily. This experiment will also show the efficiency of the approach, which makes it capable to handle large-scale dataset.

There are four reasons to choose this dataset. Firstly and most importantly, this dataset has user created object segmentation and detailed labeling information. The segmentation and labeling information can be as detail as parts of the objects, such as the front light of the car, the door of a building, etc. This detailed labeling information is considered as the supervised information to learn the distance metric. Secondly, it contains around 11,281 objects of 495 categories, which are frequently appear in daily life. It is a great challenging for any model to detect and annotate these objects in such a complex situation. As far as I know, there are few experiments conducted on such a complex object dataset in computer vision literature. Thirdly, the number of training images in this dataset is not large. Generally, it is difficult to collect large number of images with user segmentations and labels about an object. It is a challenging problem to learn an effective metric on limited supervised information. Fourthly, all the images are high resolution and generated from real world. Images in this dataset have the same resolution as photos from ordinary digital cameras.

6.2 Image Representation

In this experiment, we adopt the SIFT feature to represent the objects. There are three reasons to use the SIFT features. Firstly, the SIFT feature is believed invariant for object scaling, rotation and affine invariance changes, which is relatively more robust than other features, especially for topic of objects. Secondly, the SIFT feature performs well on street scenarios. Since the testbed contains around 11,281 objects of 495 categories related to street views, the adoption of the SIFT feature makes sense. Finally, the traditional bag of words model uses the SIFT feature, it is reasonable to adopt the same type of feature in the comparison.

6.3 Experimental Settings

In this experiment, we compare the proposed SPBoW with original BoW model to show the effectiveness of the codebook. We also aim to evaluate and compare different metric learning approaches with the proposed learning algorithm for the codebook generation.

For the traditional bag of words model, the codebook is generated by k-means over all the visual features in the dataset. The center of the clusters is deemed as the location of the code in the feature space, and each cluster is indexed by a code. All the codes and their locations in feature space form the codebook. Then each feature is assigned to the nearest code in the feature space. In this way a histogram of the codes can be generated for each image. If the dimensions of the histogram related to certain objects are high, the annotation of the object is added to the image.

To examine how existing distance metric learning methods can be beneficial to SPBoW, we implement the SPBoW methods by adapting another four of the state of the arts metric learning algorithms, including RCA [1], Information Theoretic Metric Learning algorithm (ITML)[8], Large Margin Nearest Neighbor (LMNN)[35], Neighborhood Components Analysis (NCA)[13], in the same experiment settings.

Similar to RCA, SPBoW also tries to learn a distance metric to map the related features closer. The difference is that the RCA method puts all the features on the same object into one chunklet, and SPBoW assumes the same object is consists of several semantics (i.e. car is consists of window, door, tyre, lights, etc.), and maps the features of the same objects into multiple codes to minimized the distance between the features with same semantic meaning. Firstly, the features related to each of the objects is extracted from the dataset with the help of the object location information in Labelme data. Then a collection of feature pairs are sampled randomly from the dataset. If the pair of features are on the same type of object, marked 1, unknown 0, otherwise -1. These marked feature pairs are used to train a distance metric for the proposed method. For each type of object, a codebook is generated by clustering all the features with the object. Each cluster will form a code, and the largest distance between two samples in the cluster is defined as the range of the code. The features within the range will be assigned to this code. Codebook of all the objects are combined together to form a global codebook, and thus we know the correspondence between the codes and objects. For each image, the low level features is assigned to multiple codebooks. (each feature can be assigned to more than one code if it lies in the range of multiple codes.) A histogram over the global codebook is generated.

6.4 Experiment I: Annotation Performance

In this experiment, we evaluate the annotation performance of these methods by adopting *Average Precision* (AP@N) and *Average Recall* (AR@N) in the image annotation task. Given an image, the task is to annotate the image with tags representing some semantic objects. The ground truth is generated by web users from Labelme project [29]. To evaluate the annotation performance, we adopt average precision and recall at top N (ranging from 10 to 100) annotated tags.

$$AP@N = \frac{\text{Average Number of Correct Tags in Top N}}{\text{N}} \quad (13)$$

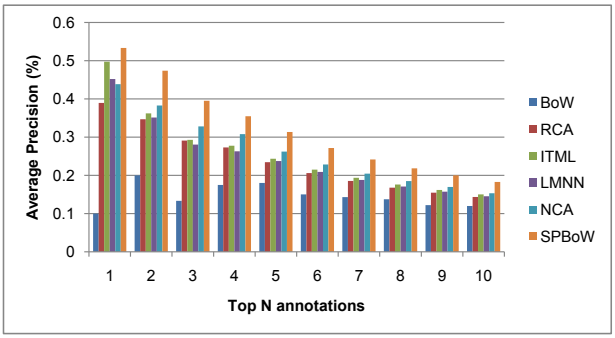
$$AR@N = \frac{\text{Average Number of Correct Tags in Top N}}{\text{Total Number of Tags in Groundtruth}} \quad (14)$$

We adopt 5-fold cross validation to run the experiment, in which 4 folds are used for building the codebook and 1 fold is used for testing annotation performance. There are two main parameters: the number of pair-wise constraints and the codebook size. In this experiment, we simply fix the constraint size to 10,000 and the codebook size to 2500. We will evaluate their influences in latter experiments. Fig. 2 shows the comparison results of different approaches.

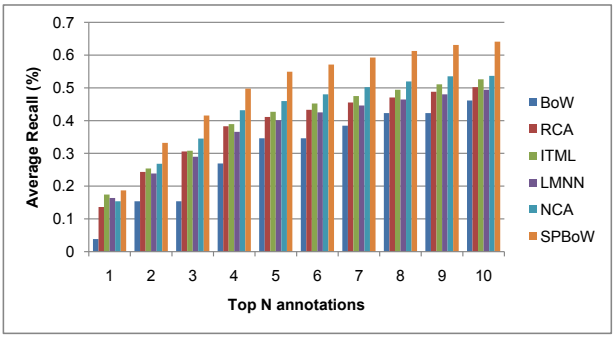
From Fig. 2, we found the SPBoW method has significantly improves the performance of the traditional BOW method in both annotation precision and recall. Comparing with other existing metric learning algorithms, SPBoW with the new metric learning algorithm also shows its advantage. This result also shows that the codebook generated by SPBoW is more discriminative than the traditional BoW method, and SPBoW is effective in learning a semantic metric between low level appearance features.

6.5 Experiment II: Varied Codebook Sizes

In this section, we evaluate the influence of the codebook size on the final annotation performance. We gradually increase the size of the codebook from 2,500 to 4,500, and record the average precision under each codebook setting.



(a) Average Precision



(b) Average Recall

Figure 2: Model comparison

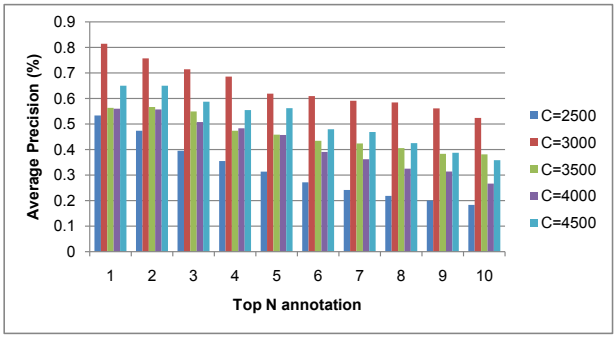


Figure 3: AP@N (from 1 to 10) under each codebook size. C represents the size of the codebook.

Fig.3 shows the size of codebook can greatly influence the performance of final annotation, with the increase of the size, the average precision at top N annotations will first increase and then fall. The empirically optimal codebook size for this dataset is at 3,000.

6.6 Experiment III: Time Cost Evaluation

In this experiment, we evaluate the time cost performance of these methods. In our approach, as we use 5-fold cross validation approach, 4 folds of the data are used to generate the codebook and 1 fold used for object annotation. We focus on measuring the computational time on codebook generation by the methods. The annotation time costs are almost similar for all the compared methods.

Table 1 shows the average computational time of generating the codebook including the metric learning time. For the BoW model, the codebook of 2,500 visual words is generated by k-means over 50,000 random features sampled from all objects. For the other methods, codebooks are generated for each object category and then combined. From the comparison, we find RCA and SPBoW are efficient in the codebook generation process, even faster than the BoW method. This is due to both the efficient learning algorithm and the clustering scheme. The comparison between different metric learning algorithms shows the efficiency of the proposed algorithm. The comparison with the BoW method, which does not need metric learning, shows clustering features in each category and then combined is much more efficient than clustering over all categories.

(s)	BoW	$\text{SPBoW}^{\text{RCA}}$	$\text{SPBoW}^{\text{ITML}}$
Codebook	121	3	96
(s)	$\text{SPBoW}^{\text{LMNN}}$	$\text{SPBoW}^{\text{NCA}}$	SPBoW
Codebook	1759	457	8

Table 1: Time cost comparisons of codebook generation processes.

7. CONCLUSION

We proposed a novel codebook generation method for the bag of words model, in which different categories of objects in images are segmented by users and the features on the same object are collected together to learn a semantics preserving codebook. We also propose a metric learning algorithm to learn the metric by minimizing the semantic gap. We formulate the task as a quadratic programming problem, and solve it with an efficient eigin vector projection algorithm. The learnt semantics preserving codebook can be used to generate the SPBoW model of the objects. Experiments on 11,281 objects from 495 categories show that the proposed metric learning algorithm is effective and outperforms other metric learning methods in object annotation. The learnt codebook significantly improves the performance of the BoW model in object annotation task.

Acknowledgments

The work was supported in part by Singapore MOE Academic Tier-1 Grant (RG67/07), the National High Technology Research and Development Program of China (863) (2008AA01Z117), the Research Fund for the Doctoral Program of Higher Education (20070358040), and National Natural Science Foundation of China (60672056).

8. REFERENCES

[1] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning distance functions using equivalence relations. In *Proc. of ICML*, pages 11–18, 2003.

[2] K. Barnard, P. Duygulu, D. Forsyth, N. D. Freitas, D. M. Blei, J. K, T. Hofmann, T. Poggio, and J. Shawe-taylor. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[3] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:2003, 2003.

[4] L. Cao and L. Fei-Fei. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *Proc. of ICCV*, pages 1–8, 2007.

[5] G. Carneiro and N. Vasconcelos. Formulating semantic image annotation as a supervised learning problem. In *Proc. of CVPR*, pages 163–168, 2005.

[6] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[7] C. Dance, J. Willamowski, L. Fan, C. Bray, and G. Csurka. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, 2004.

[8] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *Proc. of ICML*, pages 209–216, 2007.

[9] P. Duygulu, K. Barnard, J. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proc. of ECCV*, pages 97–112, 2002.

[10] J. Fan, Y. Gao, and H. Luo. Multi-level annotation of natural scenes using dominant image components and semantic concepts. In *ACM Multimedia*, pages 540–547, 2004.

[11] A. Globerson and S. Roweis. Metric learning by collapsing classes. In *Proc. of NIPS*, 2005.

[12] K.-S. Goh, B. Li, and E. Chang. Using one-class and two-class svms for multiclass image annotation. *IEEE Trans. on Knowl. and Data Eng.*, 17(10):1333–1346, 2005.

[13] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighborhood component analysis. In *Proc. of NIPS*, 2004.

[14] V. Hedau, H. Arora, and N. Ahuja. Matching images under unstable segmentations. In *Proc. of CVPR*, pages 1–8, 2008.

[15] T. Hofmann. Probabilistic latent semantic indexing. In *Proc. of SIGIR*, pages 50–57, 1999.

[16] C.-H. Hoi and M. R. Lyu. A novel log-based relevance feedback technique in content-based image retrieval. In *Proc.ACM Multimedia Conference, New York*, 2004.

[17] S. C. H. Hoi, W. Liu, M. R. Lyu, and W.-Y. Ma. Learning distance metrics with contextual constraints for image retrieval. In *Proc. of CVPR*, 2006.

[18] S.C.H. Hoi, W. Liu, and S.-F. Chang. Semi-Supervised Distance Metric Learning for Collaborative Image Retrieval. In *Proc. of CVPR*, 2008.

[19] Y.-G. Jiang, C.-W. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *Proc. of CIVR*, pages 494–501, 2007.

[20] R. Jin, J. Y. Chai, and L. Si. Effective automatic image annotation via a coherent language model and active learning. In *ACM Multimedia*, pages 892–899, 2004.

[21] C. H. Lampert, M. B. Blaschko, and T. Hofmann. Beyond sliding windows: object localization by efficient subwindow search. In *Proc. of CVPR*, 2008.

[22] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. of CVPR*, 2006.

[23] L. Fei-Fei, and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proc. of CVPR*, pages 524–531, 2005.

[24] J. Li and J. Z. Wang. Real-time computerized annotation of pictures. In *ACM Multimedia*, 2006.

[25] J. Li, W. Wu, T. Wang, and Y. Zhang. One step beyond histograms: Image representation using markov stationary features. In *Proc. of CVPR*, pages 1–8, 2008.

[26] D. G. Lowe. Object recognition from local scale-invariant features. In *Proc. of ICCV*, pages 1150–1157, 1999.

[27] D. G. Lowe. Distinctive image features from scale-invariant keypoints. In *Proc. of IJCV*, 60:91–110, 2004.

[28] R. Maree, P. Geurts, J. Piater, and L. Wehenkel. Random subwindows for robust image classification. In *Proc. of CVPR*, pages 34–40, 2005.

[29] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation. *Int. J. Comput. Vision*, 77(1-3):157–173, 2008.

[30] L. Si, R. Jin, S. C. H. Hoi, and M. R. Lyu. Collaborative image retrieval via regularized metric learning. *ACM Multimedia Systems Journal*, 12(1):34–44, 2006.

[31] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering object categories in image collections. In *Proc. of ICCV*, 2005.

[32] P. Tirilly, V. Claveau, and P. Gros. Language modeling for bag-of-visual words image categorization. In *Proc. of CIVR*, pages 249–258, 2008.

[33] E. Tola, V. Lepetit, and P. Fua. A fast local descriptor for dense matching. In *Proc. of CVPR*, pages 1–8, 2008.

[34] X.-J. Wang, L. Zhang, F. Jing, and W.-Y. Ma. Annosearch: Image auto-annotation by search. In *Proc. of CVPR*, pages 1483–1490, 2006.

[35] K. Weinberger, J. Blitzer, and L. Saul. Distance metric learning for large margin nearest neighbor classification. *Advances in Neural Information Processing Systems*, 18:1473–1480, 2006.

[36] L. Wu, Y. Hu, M. Li, N. Yu, and X.-S. Hua. Scale-invariant visual language modeling for object categorization. *Multimedia, IEEE Transactions on*, 11(2):286–294, 2009.

[37] L. Wu, X.-S. Hua, N. Yu, W.-Y. Ma, and S. Li. Flickr distance. In *ACM Multimedia*, pages 31–40, 2008.

[38] L. Wu, M. Li, Z. Li, W.-Y. Ma, and N. Yu. Visual language modeling for image classification. In *ACM Multimedia Workshop on multimedia information retrieval*, pages 115–124, 2007.

[39] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning with application to clustering with side-information. In *Proc. of NIPS*, 2002.

[40] L. Yang, R. Jin, R. Sukthankar, and Y. Liu. An efficient algorithm for local distance metric learning. In *Proc. of AAAI*, 2006.

[41] L. Yang, R. Jin, R. Sukthankar, and F. Jurie. Unifying discriminative visual codebook generation with classifier training for object category recognition. In *Proc. of CVPR*, pages 1–8, 2008.